EDA / Descriptive Statistics

## Introduction:

NBA Basketball is one of the most followed game in US and across the world. NBA which is a premier profession basketball league in which 30 teams participates and each team plays around 82 games every year generates huge amount of data. Good visualization is important to get clear insights from this vast amount of data. Our visualization project will contain dashboards which will give insights into performance of the teams and players over a period of time.

## Overall design strategy

The data which was extracted to get insights into team and player performance consisted of more than 1.04 million records which covers information related games played from around 1965. This data is displayed in two dashboards to give insights.

As the data is present at a player game level we had to create a separate data connection to aggregate the data at team level for few of the visualizations. Overall we have used both the raw data and a custom SQL output to run the visualization.

After the data is prepared, visualization color codes are identified to show the win and loss metrics for each of the dashboard. Overall we have use Green for win and Red for game lost in each of the worksheet.

For consistency purpose Trebuchet MS font is used for all the titles and subtitles in the dashboard. For all numerical axis labels Cambria Math font with 8 size is used and for textual axis labels Calibri 8 font is used

For dashboard, background color light blue is used and for all titles background color of light green is used. For all tool tips and titles, where ever emphasis was required to highlight the data different colors are used.

## Data Overview

Data is extracted from http://www.basketball-reference.com/ site which contains NBA basketball statistics and historical data. Player by Player data will be collected to use in the visualization. Around 1.04 million records are extracted for player game level data and later awards, championships, player demographic data is extracted. All these reference data is less than 200 records except player demographic which is around 4200 records.

Data is extracted into text files and when uploading into tableau all these files are joined based on the keys. Once the data is loaded another extract summarizing the team level data is created using custom SQL.

## Users

Users of the visualization dashboard will be:

* Team Coaches: Team coaches can use these visualizations to get insights into their team performance, their strengths and weakness and use that to build the team accordingly
* Player: To understand how they are performing over a period of time
* Fans: To get a get a quick information on their favorite team or player

## Questions

Questions which will be answered by this visualization:

**Team Coach:**

* What is the trend in the performance of the team
* Which are worst lost games and best win games of the season
* Against which team the team is strong and against which they are weak
* How is the performance varying when played in home ground and non-home grounds

**Players:**

* How are they performing from game to game in a season
* What is the shooting success rate
* Against which teams they perform well and against whom they perform bad

## Describe Visualization and how it answers the questions

**Teams:**

* What is the trend in the performance of the team?
  + The team game data visualization give the information and game by game points scored and whether they have won the game or lost the game
  + This visualization also gives the information on how many rebounds (overall team), assists, steals and fouls the team has made (we need to select the parameter measure to get each of the metric in the same line graph)
  + The data is also compared against the season average to get the insight on where they stand compared to other teams played in the season
* Which are worst lost games and best win games of the season
  + There are two graphs which display the top 5 best win and top 5 worst lost games. This visualization will give the information on by how many points they have won or lost and against which teams. This can help in preparing the strategy when playing with these teams and can improve
* Against which team the team is strong and against which they are weak
  + The graph which is prepared here will give the information on the % of win and loss against each of the teams they have played and the graph is prepared to dynamically sort by winning% to get a quick view on against which team they are strong and against whom they are weak.
* How is the performance varying when played in home ground and non-home grounds
  + This bar graph gives the information on the % of win and % of loss when the team plays in their own team ground and when played at opponent teams ground. This will give information to see if the team is weak in non-home ground games.

**Players:**

* How are the performing from game to game in a season
  + This visualization gives information on the trends in player’s games in a particular season. This visualization also gives the information on how may rebounds they made, steals, assists and fouls in a particular game of that season. (we need to select the parameter measure to get each of the metric in the same line graph)
  + This graph also compares against the season’s average to get insights on where the player stands compared to rest of the players who played in that season
* What is the shooting success rate
  + This bar graph gives the information on the % of success in each of the field goals, 3 pointers and free throws
* Against which teams they perform well and against whom they perform bad
  + This bar graph which contains information related to points, rebounds, steals, assists, fouls is compared against each of the opponent team the player has played. This gives the information on against which team the player is doing good and against who he is not doing well.

## Conclusion

It is very difficult to analyze the NBA data which is more than 1million records even after aggregating the information. Aggregated numerical data is good to some extent but doesn’t give the connecting information and doesn’t give the flexibility of multiple view of the data. Visualizing this data with various parameters will give good insights quickly and can help taking actions to improve the performance.

With the latest player tracking technology where each of the players movements are tracked, we will have a good opportunity to build visualization to analyze ever further and get more insights on players performance.